

# Signal Processing based on Stable radix-2 DCT Algorithms having Orthogonal Factors

Sirani M. Perera

**Keywords:** Orthogonal DCT Factorization, Recursive, Stable radix-2 DCT Algorithms, Error Bounds, Image Compression, Signal Flow Graphs

## Abstract

This paper presents stable, radix-2, completely recursive discrete cosine transformation algorithms DCT-I and DCT-III solely based on DCT-I, DCT-II, DCT-III, and DCT-IV having sparse and orthogonal factors. Error bounds for computing the completely recursive DCT-I, DCT-II, DCT-III, and DCT-IV algorithms having sparse and orthogonal factors are addressed. Image compression results are presented based on the recursive 2D DCT-II and DCT-IV algorithms for image size  $512 \times 512$  pixels with transfer block sizes  $8 \times 8$ ,  $16 \times 16$ , and  $32 \times 32$  with 93.75% absence of coefficients in each transfer block. Finally signal flow graphs are demonstrated based on the completely recursive DCT-I, DCT-II, DCT-III, and DCT-IV algorithms having orthogonal factors.

## 1 INTRODUCTION

The Fast Fourier Transform is used to efficiently compute the Discrete Fourier Transform (DFT) and its inverse. The DFTs are widely used in numerous applications in applied mathematics and electrical engineering [27, 23, 24, 3, 19, 30], etc.

The DFT uses complex arithmetic. The DFT of a sequence of  $n$ -input  $\{x_k\}_{k=0}^{n-1}$  is the sequence of  $n$ -output  $\{y_k\}_{k=0}^{n-1}$  defined via

$$\begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_{n-1} \end{bmatrix} = \frac{1}{\sqrt{n}} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & \omega_n & \cdots & \omega_n^{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \omega_n^{n-1} & \cdots & \omega_n^{(n-1)(n-1)} \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{n-1} \end{bmatrix} \quad (1)$$

where  $\omega_n = e^{-\frac{2\pi i}{n}}$ . There exist real analogues of the DFT, namely the Discrete Cosine Transforms and Discrete Sine Transforms, the main types are from I to IV. Similar to (1), the I-IV variants of cosine and sine matrices transform the sequence of  $n$ -input into a sequence of  $n$ -output via the transform matrices stated in Table (1), where for DCT-I  $j, k = 0, 1, \dots, n$ , DST-I  $j, k = 0, 1, \dots, n-2$ , DCT and DST II-IV  $j, k = 0, 1, \dots, n-1$ ,  $\epsilon_n(0) = \epsilon_n(n) = \frac{1}{\sqrt{2}}$ ,  $\epsilon_n(j) = 1$  for  $j \in \{1, 2, \dots, n-1\}$  and  $n \geq 2$  is an integer. Among DCT I-IV transformations,  $C_{n+1}^I$  was introduced in [31],  $C_n^{II}$  and its

Cosine and Sine Transforms	Inverse Transforms
$C_{n+1}^I = \sqrt{\frac{2}{n}} \left[ \epsilon_n(j) \epsilon_n(k) \cos \frac{jk\pi}{n} \right]$	$[C_{n+1}^I]^{-1} = C_{n+1}^I$
$C_n^{II} = \sqrt{\frac{2}{n}} \left[ \epsilon_n(j) \cos \frac{j(2k+1)\pi}{2n} \right]$	$[C_n^{II}]^{-1} = C_n^{III}$
$C_n^{III} = \sqrt{\frac{2}{n}} \left[ \epsilon_n(k) \cos \frac{(2j+1)k\pi}{2n} \right]$	$[C_n^{III}]^{-1} = C_n^I$
$C_n^{IV} = \sqrt{\frac{2}{n}} \left[ \cos \frac{(2j+1)(2k+1)\pi}{4n} \right]$	$[C_n^{IV}]^{-1} = C_n^{IV}$
$S_{n-1}^I = \sqrt{\frac{2}{n}} \left[ \sin \frac{(j+1)(k+1)\pi}{n} \right]$	$[S_{n-1}^I]^{-1} = S_{n-1}^I$
$S_n^{II} = \sqrt{\frac{2}{n}} \left[ \epsilon_n(j+1) \sin \frac{(j+1)(2k+1)\pi}{2n} \right]$	$[S_n^{II}]^{-1} = S_n^{III}$
$S_n^{III} = \sqrt{\frac{2}{n}} \left[ \epsilon_n(k+1) \sin \frac{(2j+1)(k+1)\pi}{2n} \right]$	$[S_n^{III}]^{-1} = S_n^I$
$S_n^{IV} = \sqrt{\frac{2}{n}} \left[ \sin \frac{(2j+1)(2k+1)\pi}{4n} \right]$	$[S_n^{IV}]^{-1} = S_n^{IV}$

Table 1: Cosine and Sine Transform Matrices

inverse  $C_n^{III}$  were introduced in [1], and  $C_n^{IV}$  was introduced into digital signal processing in [9]. Moreover, among DST I-IV transformations,  $S_{n-1}^I$  and  $S_n^{IV}$  were introduced in [10, 9] and  $S_n^{II}$  and its inverse  $S_n^{III}$  were introduced in [14]. These classifications were also stated in [30, 19].

It has been stated, in e.g. [21, 22, 24], that these cosine and sine matrices of types I-IV are orthogonal. Strang, in [24], proved that the column vectors of each cosine matrix are eigenvectors of a symmetric second difference matrix under different boundary conditions, and are hence orthogonal. Later Britanak, Yip, and Rao in [3] followed very closely the presentation made by Strang's [24] to point out that the column vectors of each cosine and sine matrix of types I-VIII are eigenvectors of a symmetric second difference matrix. Due to properties of these DCT and DST, it was shown by many authors (see e.g. [3, 2, 4, 6, 7, 11, 14, 13, 12, 15, 16, 17, 24, 28, 29]) that these symmetric and asymmetric (rarely used) versions of DCT and DST can be widely used in image processing, signal processing, finger print enhancement, quick response code (QR code), etc.

To obtain real, fast DCT or DST algorithms one can mainly use a polynomial arithmetic technique or a matrix factorization technique. In the polynomial arithmetic technique (see e.g. [25]), components of  $C_n \mathbf{x}$  or  $S_n \mathbf{x}$  are

interpreted as the nodes of a degree  $n$  polynomial, and then one applies the divide and conquer technique to reduce the degree of the polynomial. Later it was found (see e.g. [26]) that the polynomial arithmetic technique leads to inferior numerical stability of the DCT and DST algorithms. The matrix factorization technique is the direct factorization of the DCT or DST matrices into the product of sparse matrices (see e.g. [30, 32, 3, 19, 18]). The matrix factorization for DST-I in [32] used the results in [5] to decompose DST-I into DCT and DST. Also the decomposition for DCT-II in [30] is a slightly different version of the result in [5]. Though one can find orthogonal matrix factorizations for DCT and DST in [30], the resulting algorithms in [30] are not completely recursive, and hence do not lead to simple recursive algorithms. Moreover [3] has used the same factorization for DST-II and DST-IV as in [30]. On the other hand, one can use these [30, 3, 24] results to derive recursive, stable algorithms as stated in [19, 18].

However, [19] has offered stable, recursive DCT-II and DCT-IV algorithms, based on DCT-II and DCT-IV. Thus this paper completes the picture and provides completely recursive, stable, radix-2 DCT-I and DCT-III algorithms that are solely defined via DCT I-IV, having sparse and orthogonal factors. The paper also addresses the error bounds on computing completely recursive algorithms for DCT I-IV. Moreover, this paper elaborates image compression (absence of 93.75% coefficients in each transfer block) and signal transform designs based on the completely recursive algorithms based on DCT I-IV.

In section 2 we derive factorizations for DCT-I and DCT-III having orthogonal and sparse matrices, and state completely recursive DCT I-IV algorithms solely defined via DCT I-IV having sparse, orthogonal, and rotation/rotation-reflection matrices. Next, in section 3, we present the arithmetic cost of computing these algorithms. In section 4 we derive error bounds in computing these algorithms and discuss the stability. Finally in sections 5 and 6 respectively, we demonstrate image compression results and signal flow graphs based on these completely recursive DCT I-IV algorithms.

## 2 COMPLETELY RECURSIVE RADIX-2 DCT ALGORITHMS HAVING ORTHOGONAL FACTORS

This section introduces sparse and orthogonal factorizations for DCT-I and DCT-III matrices. In the meantime, we present completely recursive, radix-2 DCT I-IV algorithms solely defined via DCT I-IV, having sparse, orthogonal, and butterfly matrices. One can observe a variant of the DCT-II and DCT-IV algorithms having almost orthogonal factors in [19].

The following notations and sparse matrices are used frequently in this paper. Denote an involution matrix  $\tilde{I}_n$  by  $\tilde{I}_n \mathbf{x} = [x_{n-1}, x_{n-2}, \dots, x_0]^T$ , a diagonal matrix  $D_n$  by  $D_n \mathbf{x} = \text{diag}((-1)^k)_{k=0}^{n-1} \mathbf{x}$ , an even-odd permutation matrix  $P_n$  ( $n \geq 3$ ) by

$$P_n \mathbf{x} = \begin{cases} [x_0, x_2, \dots, x_{n-2}, x_1, x_3, \dots, x_{n-1}]^T & \text{even } n, \\ [x_0, x_2, \dots, x_{n-1}, x_1, x_3, \dots, x_{n-2}]^T & \text{odd } n, \end{cases}$$

for any  $\mathbf{x} = [x_j]_{j=0}^{n-1}$ , and orthogonal matrices ( $n \geq 4$ ) by

$$\check{H}_{n+1} = \frac{1}{\sqrt{2}} \begin{bmatrix} I_{\frac{n}{2}} & \tilde{I}_{\frac{n}{2}} \\ I_{\frac{n}{2}} & -\tilde{I}_{\frac{n}{2}} \end{bmatrix} \sqrt{2}, \quad H_n = \frac{1}{\sqrt{2}} \begin{bmatrix} I_{\frac{n}{2}} & \tilde{I}_{\frac{n}{2}} \\ I_{\frac{n}{2}} & -\tilde{I}_{\frac{n}{2}} \end{bmatrix},$$

$$U_n = \begin{bmatrix} 1 & & \\ & \frac{1}{\sqrt{2}} \begin{bmatrix} I_{\frac{n}{2}-1} & I_{\frac{n}{2}-1} \\ I_{\frac{n}{2}-1} & -I_{\frac{n}{2}-1} \end{bmatrix} & \\ & & -1 \end{bmatrix} \begin{bmatrix} I_{\frac{n}{2}} & \\ & D_{\frac{n}{2}} \tilde{I}_{\frac{n}{2}} \end{bmatrix},$$

$$R_n = \begin{bmatrix} I_{\frac{n}{2}} & \\ & D_{\frac{n}{2}} \end{bmatrix} \begin{bmatrix} \text{diag } C_{\frac{n}{2}} & (\text{diag } S_{\frac{n}{2}}) \tilde{I}_{\frac{n}{2}} \\ -\tilde{I}_{\frac{n}{2}} (\text{diag } S_{\frac{n}{2}}) & \text{diag } (\tilde{I}_{\frac{n}{2}} C_{\frac{n}{2}}) \end{bmatrix}$$

where for  $k = 0, 1, \dots, \frac{n}{2} - 1$

$$C_{\frac{n}{2}} = \left[ \cos \frac{(2k+1)\pi}{4n} \right] \quad \text{and} \quad S_{\frac{n}{2}} = \left[ \sin \frac{(2k+1)\pi}{4n} \right].$$

DCT-II and DCT-IV algorithms are the keys for the completely recursive procedure, so for a given vector  $\mathbf{x} \in \mathbb{R}^n$ , we present algorithms in order  $\mathbf{y} = C_n^{II} \mathbf{x}$ ,  $\mathbf{y} = C_n^{IV} \mathbf{x}$ ,  $\mathbf{y} = C_n^{III} \mathbf{x}$  and  $\mathbf{y} = C_{n+1}^I \mathbf{x}$ . Following the matrix factorizations for DCT-II and DCT-IV in [19], let us first state recursive DCT-II and DCT-IV having orthogonal factors via algorithms (2.1) and (2.2), respectively.

### Algorithm 2.1. ( $\cos 2(\mathbf{x}, \mathbf{n})$ )

*Input:*  $n = 2^t$  ( $t \geq 1$ ),  $n_1 = \frac{n}{2}$ ,  $\mathbf{x} \in \mathbb{R}^n$ .

1. If  $n = 2$ , then

$$\mathbf{y} := \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \mathbf{x}.$$

2. If  $n \geq 4$ , then

$$[u_j]_{j=0}^{n-1} := H_n \mathbf{x},$$

$$\mathbf{z1} := \cos 2 \left( [u_j]_{j=0}^{n_1-1}, n_1 \right),$$

$$\mathbf{z2} := \cos 4 \left( [u_j]_{j=n_1}^{n-1}, n_1 \right),$$

$$\mathbf{y} := P_n^T (\mathbf{z1}^T, \mathbf{z2}^T)^T.$$

*Output:*  $\mathbf{y} = C_n^{II} \mathbf{x}$ .

**Algorithm 2.2.**  $(\cos 4(\mathbf{x}, \mathbf{n}))$ 

Input:  $n = 2^t (t \geq 1)$ ,  $n_1 = \frac{n}{2}$ ,  $\mathbf{x} \in \mathbb{R}^n$ .

1. If  $n = 2$ , then

$$\mathbf{y} := \begin{bmatrix} \cos \frac{\pi}{8} & \sin \frac{\pi}{8} \\ \sin \frac{\pi}{8} & -\cos \frac{\pi}{8} \end{bmatrix} \mathbf{x}.$$

2. If  $n \geq 4$ , then

$$[u_j]_{j=0}^{n-1} := R_n \mathbf{x},$$

$$\mathbf{z1} := \cos 2 \left( [u_j]_{j=0}^{n_1-1}, n_1 \right),$$

$$\mathbf{z2} := \cos 2 \left( [u_j]_{j=n_1}^{n-1}, n_1 \right),$$

$$\mathbf{w} := U_n (\mathbf{z1}^T, \mathbf{z2}^T)^T,$$

$$\mathbf{y} := P_n^T \mathbf{w}.$$

Output:  $\mathbf{y} = C_n^{IV} \mathbf{x}$ .

By using the well known transpose property between DCT-II and DCT-III we can state an algorithm for DCT-III via (2.3). This algorithm executes recursively with the DCT-II and DCT-IV algorithms.

**Algorithm 2.3.**  $(\cos 3(\mathbf{x}, \mathbf{n}))$ 

Input:  $n = 2^t (t \geq 1)$ ,  $n_1 = \frac{n}{2}$ ,  $\mathbf{x} \in \mathbb{R}^n$ .

1. If  $n = 2$ , then

$$\mathbf{y} := \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \mathbf{x}.$$

2. If  $n \geq 4$ , then

$$[u_j]_{j=0}^{n-1} := P_n \mathbf{x},$$

$$\mathbf{z1} := \cos 3 \left( [u_j]_{j=0}^{n_1-1}, n_1 \right),$$

$$\mathbf{z2} := \cos 4 \left( [u_j]_{j=n_1}^{n-1}, n_1 \right),$$

$$\mathbf{y} := H_n^T (\mathbf{z1}^T, \mathbf{z2}^T)^T.$$

Output:  $\mathbf{y} = C_n^{III} \mathbf{x}$ .

Before stating the algorithm for DCT-I let us derive a sparse and orthogonal factorization for DCT-I.

**Lemma 2.4.** Let  $n \geq 4$  be an even integer. The matrix  $C_{n+1}^I$  can be factored in the form

$$C_{n+1}^I = P_{n+1}^T \left[ \begin{array}{c|c} C_{\frac{n}{2}+1}^I & 0 \\ \hline 0 & C_{\frac{n}{2}}^{III} \end{array} \right] \check{H}_{n+1}. \quad (2)$$

*Proof.* Let's apply  $P_{n+1}$  to  $C_{n+1}^I$  to permute rows and then partition the resultant matrix. So

$$(1,1) \text{ block becomes } \sqrt{\frac{2}{n}} \left[ \epsilon_n(2j) \epsilon_n(k) \cos \frac{2jk\pi}{n} \right]_{j,k=0}^{\frac{n}{2}},$$

(1,2) block becomes

$$\begin{aligned} & \sqrt{\frac{2}{n}} \left[ \epsilon_n(2j) \epsilon_n \left( \frac{n}{2} + k + 1 \right) \cos \frac{j(n+2k+2)\pi}{n} \right]_{j,k=0}^{\frac{n}{2}, \frac{n}{2}-1} \\ &= \sqrt{\frac{2}{n}} \left[ \epsilon_n(2j) \epsilon_n \left( \frac{n}{2} + k + 1 \right) \cos \frac{j(n-2k-2)\pi}{n} \right]_{j,k=0}^{\frac{n}{2}, \frac{n}{2}-1}, \end{aligned}$$

(2,1) block becomes  $\sqrt{\frac{2}{n}} \left[ \epsilon_n(k) \cos \frac{(2j+1)k\pi}{n} \right]_{j,k=0}^{\frac{n}{2}-1, \frac{n}{2}}$ ,

(2,2) block becomes

$$\begin{aligned} & \sqrt{\frac{2}{n}} \left[ \epsilon_n \left( \frac{n}{2} + k + 1 \right) \cos \frac{(2j+1)(n+2k+2)\pi}{2n} \right]_{j,k=0}^{\frac{n}{2}-1} \\ &= \sqrt{\frac{2}{n}} \left[ -\epsilon_n \left( \frac{n}{2} + k + 1 \right) \cos \frac{(2j+1)(n-2k-2)\pi}{2n} \right]_{j,k=0}^{\frac{n}{2}-1}. \end{aligned}$$

Hence

$$\begin{aligned} P_{n+1} C_{n+1}^I &= \frac{1}{\sqrt{2}} \left[ \begin{array}{c|c} C_{\frac{n}{2}+1}^I \begin{bmatrix} I_{\frac{n}{2}} & 0 \\ 0 & \sqrt{2} \end{bmatrix} & C_{\frac{n}{2}+1}^I \begin{bmatrix} \tilde{I}_{\frac{n}{2}} \\ 0 \end{bmatrix} \\ \hline C_{\frac{n}{2}}^{III} \begin{bmatrix} I_{\frac{n}{2}} & 0 \end{bmatrix} & -C_{\frac{n}{2}}^{III} \tilde{I}_{\frac{n}{2}} \end{array} \right] \\ &= \left[ \begin{array}{c|c} C_{\frac{n}{2}+1}^I & 0 \\ \hline 0 & C_{\frac{n}{2}}^{III} \end{array} \right] \check{H}_{n+1} \end{aligned}$$

□

Thus an algorithm for DCT-I can be stated via (2.5), which executes recursively with DCT II-IV algorithms.

**Algorithm 2.5.**  $(\cos 1(\mathbf{x}, \mathbf{n} + 1))$ 

Input:  $n = 2^t (t \geq 1)$ ,  $n_1 = \frac{n}{2}$ ,  $\mathbf{x} \in \mathbb{R}^{n+1}$ .

1. If  $n = 2$ , then

$$\mathbf{y} := \frac{1}{2} \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & \sqrt{2} \\ 1 & -1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 \\ 0 & \sqrt{2} & 0 \\ 1 & 0 & -1 \end{bmatrix} \mathbf{x}.$$

2. If  $n \geq 4$ , then

$$[u_j]_{j=0}^n := \check{H}_{n+1} \mathbf{x},$$

$$\mathbf{z1} := \cos 1 \left( [u_j]_{j=0}^{n_1}, n_1 + 1 \right),$$

$$\mathbf{z2} := \cos 3 \left( [u_j]_{j=n_1+1}^n, n_1 \right),$$

$$\mathbf{y} := P_{n+1}^T (\mathbf{z1}^T, \mathbf{z2}^T)^T.$$

Output:  $\mathbf{y} = C_{n+1}^I \mathbf{x}$ .

### 3 ARITHMETIC COST OF COMPUTING DCT ALGORITHMS

We first calculate the arithmetic cost of computing DCT I-IV algorithms. Let's denote the number of additions and

multiplications required to compute - say a length  $n$  DCT II algorithm:  $\mathbf{y} = C_n^H \mathbf{x}$  by  $\#a(\text{DCT-II}, n)$  and  $\#m(\text{DCT-II}, n)$ . Note that the multiplication of  $\pm 1$  and permutations are not counted. Once the cost is computed we show numerical results for the speed improvement factor of these algorithms.

### 3.1 Number of additions and multiplications in computing DCT I-IV algorithms

Here we calculate the arithmetic cost of computing the DCT I-IV algorithms in order (2.1), (2.2), (2.3) and (2.5). The cost of addition in computing DCT-II and DCT-IV algorithms is the same as in [19], but the cost of multiplication is different from [19]. The latter is because in this paper, not only DCT-I and DCT-III algorithms but also DCT-II and DCT-IV algorithms have orthogonal factors not almost orthogonal factors. Let us first derive explicitly the number of multiplications required to compute DCT-II and DCT-IV algorithms and then the arithmetic cost of DCT-III and DCT-I algorithms respectively.

**Lemma 3.1.** *Let  $n = 2^t$  ( $t \geq 2$ ) be given. Using algorithms (2.1) and (2.2), the arithmetic cost of computing length  $n$  DCT-II algorithm is given by*

$$\begin{aligned}\#a(\text{DCT-II}, n) &= \frac{4}{3}nt - \frac{8}{9}n - \frac{1}{9}(-1)^t + 1, \\ \#m(\text{DCT-II}, n) &= \frac{5}{3}nt - \frac{10}{9}n + \frac{1}{9}(-1)^t + 1,\end{aligned}\quad (3)$$

*Proof.* Following algorithms (2.1) and (2.2)

$$\begin{aligned}\#m(\text{DCT-II}, n) &= \#m\left(\text{DCT-II}, \frac{n}{2}\right) + \#m\left(\text{DCT-IV}, \frac{n}{2}\right) \\ &\quad + \#m(H_n), \\ \#m(\text{DCT-IV}, n) &= \#m(U_n) + 2 \cdot \#m\left(\text{DCT-II}, \frac{n}{2}\right) \\ &\quad + \#m(R_n).\end{aligned}\quad (4)$$

By referring to the structures of  $H_n$ ,  $U_n$ , and  $R_n$

$$\begin{aligned}\#a(H_n) &= n, \#m(H_n) = n, \\ \#a(U_n) &= n - 2, \#m(U_n) = n - 2, \\ \#a(R_n) &= n, \#m(R_n) = 2n,\end{aligned}\quad (5)$$

Thus

$$\begin{aligned}\#m(\text{DCT-II}, n) &= \#m\left(\text{DCT-II}, \frac{n}{2}\right) + 2 \cdot \#m\left(\text{DCT-II}, \frac{n}{4}\right) \\ &\quad + \frac{5}{2}n - 2.\end{aligned}$$

Since  $n = 2^t$  we can obtain the linear difference equation of order 2 with respect to  $t$

$$\begin{aligned}\#m(\text{DCT-II}, 2^t) &- \#m(\text{DCT-II}, 2^{t-1}) - 2 \cdot \#m(\text{DCT-II}, 2^{t-2}) \\ &= 5 \cdot 2^{t-1} - 2.\end{aligned}$$

If  $\#m(\text{DCT-II}, 2^t) = \alpha^t$  (where  $\alpha \neq 0$ ) is a solution then the above follows

$$\alpha^t - \alpha^{t-1} - 2(\alpha^{t-2}) = 5 \cdot 2^{t-1} - 2. \quad (6)$$

The homogeneous solution of the above is given by solving the characteristic equation

$$\alpha^{t-2}(\alpha^2 - \alpha - 2) = 0.$$

From which we get

$$\#m(\text{DCT-II}, 2^t) = r_1 2^t + r_2 (-1)^t + \text{particular solution}$$

where  $r_1$  and  $r_2$  are constants. Let  $\alpha^t = r_3 + r_4 t \cdot 2^t$  (where  $r_3$  and  $r_4$  are constants) be the particular solution. Substituting this potential equation into (6) and equating the coefficients we can find that

$$\#m(\text{DCT-II}, 2^t) = r_1 2^t + r_2 (-1)^t + \frac{5}{3} \cdot t \cdot 2^t + 1$$

Using the initial conditions  $\#m(\text{DCT-II}, 2) = 2$  and  $\#m(\text{DCT-II}, 4) = 10$ , we can determine the general solution

$$\#m(\text{DCT-II}, 2^t) = \frac{5}{3} \cdot t \cdot 2^t - \frac{10}{9} 2^t + \frac{1}{9} (-1)^t + 1 \quad (7)$$

Thus substituting  $n = 2^t$  we can obtain the number of multiplications required to compute DCT-II algorithm as stated in (3).

Again by algorithms (2.1) and (2.2) together with (5), we can state

$$\#a(\text{DCT-II}, n) = \#a\left(\text{DCT-II}, \frac{n}{2}\right) + 2 \cdot \#a\left(\text{DCT-II}, \frac{n}{4}\right) + 2n - 2.$$

Since  $n = 2^t$ , the second order linear difference equation with respect to  $t$  can be given via

$$\begin{aligned}\#a(\text{DCT-II}, 2^t) &- \#a(\text{DCT-II}, 2^{t-1}) - 2 \cdot \#a(\text{DCT-II}, 2^{t-2}) \\ &= 2^{t+1} - 2.\end{aligned}$$

As derived analogously in the cost of multiplication, we can solve the above equation under the initial conditions  $\#a(\text{DCT-II}, 2) = 2$  and  $\#a(\text{DCT-II}, 4) = 8$  to obtain

$$\#a(\text{DCT-II}, n) = \frac{4}{3}nt - \frac{8}{9}n - \frac{1}{9}(-1)^t + 1. \quad (8)$$

□

**Corollary 3.2.** *Let  $n = 2^t$  ( $t \geq 2$ ) be given. Using algorithms (2.2) and (2.1), the arithmetic cost of computing length  $n$  DCT-IV algorithm is given by*

$$\begin{aligned}\#a(\text{DCT-IV}, n) &= \frac{4}{3}nt - \frac{2}{9}n + \frac{2}{9}(-1)^t, \\ \#m(\text{DCT-IV}, n) &= \frac{5}{3}nt + \frac{2}{9}n - \frac{2}{9}(-1)^t.\end{aligned}\quad (9)$$

*Proof.* The number of multiplications required to compute DCT-IV algorithm can be found by substituting (7) at  $\frac{n}{2}(=2^{t-1})$  into the equation (4)

$$\begin{aligned} \#m(\text{DCT-IV}, n) = n - 2 &+ 2 \left( \frac{5}{3} \cdot \frac{n}{2} (t-1) - \frac{10}{9} \cdot \frac{n}{2} \right. \\ &\left. + \frac{1}{9} (-1)^{t-1} + 1 \right) + 2n \end{aligned}$$

Simplifying the above gives the cost of multiplication

$$\#m(\text{DCT-IV}, n) = \frac{5}{3}nt + \frac{2}{9}n - \frac{2}{9}(-1)^t.$$

Similarly, the number of additions required to compute DCT-IV algorithm can be found by substituting (8) at  $\frac{n}{2}(=2^{t-1})$  to

$$\begin{aligned} \#a(\text{DCT-IV}, n) &= \#a(U_n) + 2 \cdot \#a\left(\text{DCT-II}, \frac{n}{2}\right) + \#a(R_n) \\ &= 2 \cdot \#a\left(\text{DCT-II}, \frac{n}{2}\right) + 2n - 2. \end{aligned}$$

Simplifying the above yields

$$\#a(\text{DCT-IV}, n) = \frac{4}{3}nt - \frac{2}{9}n + \frac{2}{9}(-1)^t.$$

□

The DCT-III algorithm (2.3) was stated using the transpose property of matrices so the following corollary is trivial.

**Corollary 3.3.** *Let  $n = 2^t$  ( $t \geq 2$ ) be given. If DCT-III could be computed by using algorithms (2.3), (2.2), and (2.1) then the arithmetic cost of computing a length  $n$  DCT-III algorithm is given by*

$$\begin{aligned} \#a(\text{DCT-III}, n) &= \frac{4}{3}nt - \frac{8}{9}n - \frac{1}{9}(-1)^t + 1, \\ \#m(\text{DCT-III}, n) &= \frac{5}{3}nt - \frac{10}{9}n + \frac{1}{9}(-1)^t + 1. \end{aligned} \quad (10)$$

**Remark 3.4.** By using the DCT-III algorithm (2.3) and the arithmetic cost of computing the DCT-IV algorithm (in corollary (3.2)), one can obtain the same results as in corollary (3.3).

Let us state the arithmetic cost of computing the DCT-I algorithm (2.5).

**Lemma 3.5.** *Let  $n = 2^t$  ( $t \geq 2$ ) be given. Using algorithms (2.5), (2.3), (2.2) and (2.1), the arithmetic cost of a DCT-I algorithm of length  $n+1$  is given by*

$$\begin{aligned} \#a(\text{DCT-I}, n+1) &= \frac{4}{3}nt - \frac{14}{9}n + \frac{1}{18}(-1)^t + t + \frac{7}{2} \\ \#m(\text{DCT-I}, n+1) &= \frac{5}{3}nt - \frac{22}{9}n - \frac{1}{18}(-1)^t + t + \frac{11}{2} \end{aligned} \quad (11)$$

*Proof.* Referring to the DCT-I algorithm (2.5)

$$\begin{aligned} \#a(\text{DCT-I}, n+1) &= \#a\left(\text{DCT-I}, \frac{n}{2} + 1\right) + \#a\left(\text{DCT-III}, \frac{n}{2}\right) \\ &\quad + \#a(\check{H}_{n+1}) \end{aligned} \quad (12)$$

The structure of  $\check{H}_{n+1}$  leads to  $\#a(\check{H}_{n+1}) = n$ . This together with the arithmetic cost of computing DCT-III (10) algorithm, we can rewrite (12)

$$\begin{aligned} \#a(\text{DCT-I}, n+1) &= \#a\left(\text{DCT-I}, \frac{n}{2} + 1\right) + \frac{2}{3}nt - \frac{1}{9}n \\ &\quad + \frac{1}{9}(-1)^t + 1 \end{aligned}$$

Since  $n = 2^t$ , the above simplifies to the first order linear difference equation (respect to  $t \geq 2$ )

$$\begin{aligned} \#a(\text{DCT-I}, 2^t + 1) - \#a(\text{DCT-I}, 2^{t-1} + 1) &= \frac{2}{3}t \cdot 2^t - \frac{1}{9}2^t + \frac{1}{9}(-1)^t + 1 \end{aligned} \quad (13)$$

We can obtain the number of additions required to compute the DCT-I algorithm by solving (13) under the initial condition  $\#a(\text{DCT-I}, 3) = 4$ . Analogously, one can solve the first order linear difference equation under the initial condition  $\#m(\text{DCT-I}, 3) = 5$  to obtain the number of multiplications. □

## 3.2 Speed improvement factor of DCT I-IV algorithms

Based on the results in lemmas 3.1, 3.5 and corollaries 3.2, 3.3, we graph the speed improvement factor of DCT I-IV algorithms having orthogonal factors. It is known that the speed improvement factor plays a critical role in the DFT algorithms as it gives us an idea about the processing speed of the algorithms. We should recall here that this factor increases with the size of matrix.

In our case, the speed improvement factor says the ratio between the number of additions and multiplications required to compute the DCT I-IV algorithms, and the direct computation cost of computing these algorithms which is  $2n^2 - n$  for DCT II-IV, and  $2n^2 + 3n + 1$  for DCT-I. Figure 1 shows the speed improvement factor corresponding to the DCT I-IV algorithms with respect to the size of matrix. These numerical data correspond to MATLAB (R2014a version) with machine precision  $2.2 \times 10^{-16}$ .

## 4 ERROR BOUNDS AND STABILITY OF DCT ALGORITHMS

Error bounds and stability of computing the DCT I-IV algorithms are the main concern in this section. Here, to verify the stability, we will use error bounds (using perturbation of

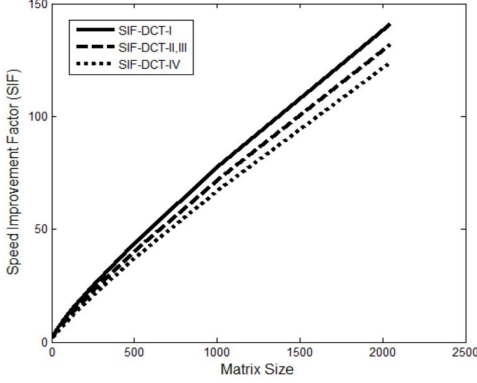


Figure 1: Speed improvement factor of DCT I-IV algorithms

the product of matrices stated in [8]) in computing these algorithms. Let us assume that the computed trigonometry functions ( $d_r := \sin \frac{r\pi}{4n}$  or  $\cos \frac{r\pi}{4n}$  are the entries of the butterfly matrix)  $\hat{d}_r$  are used and satisfy

$$\hat{d}_r = d_r + \varepsilon_r, \quad |\varepsilon_r| \leq \mu \quad (14)$$

for all  $r = 1, 3, 5, \dots, n-1$ , where  $\mu := O(u)$  and  $u$  is the unit roundoff.

Let's recall the perturbation of the product of matrices stated in [8] i.e. if  $A_k + \Delta A_k \in \mathbb{R}^{n \times n}$  satisfies  $|\Delta A_k| \leq \delta_k |A_k|$  for all  $k$ , then

$$\left| \prod_{k=0}^m (A_k + \Delta A_k) - \prod_{k=0}^m A_k \right| \leq \left( \prod_{k=0}^m (1 + \delta_k) - 1 \right) \prod_{k=0}^m |A_k|$$

where  $|\delta_k| < u$ . Moreover, recall  $\prod_{k=1}^n (1 + \delta_k)^{\pm 1} = 1 + \theta_n$  where  $|\theta_n| \leq \frac{nu}{1-nu} =: \gamma_n$  and  $\gamma_k + u \leq \gamma_{k+1}$ ,  $\gamma_k + \gamma_j + \gamma_k \gamma_j \leq \gamma_{k+j}$  from [8].

Let us derive error bounds for computing recursive DCT I-IV algorithms with the help of the perturbations in a matrix product.

**Theorem 4.1.** Let  $\hat{\mathbf{y}} = fl(C_n^{II} \mathbf{x})$ , where  $n = 2^t$  ( $t \geq 2$ ), be computed using the algorithms (2.1), (2.2), and assume that (14) holds, then

$$\frac{\|\mathbf{y} - \hat{\mathbf{y}}\|_2}{\|\mathbf{y}\|_2} \leq \frac{\gamma_7(t-1)}{1 - \gamma_7(t-1)}. \quad (15)$$

*Proof.* Using the algorithms (2.1), (2.2), and the computed matrices  $\hat{\mathbf{G}}_k$  (in terms of the computed  $\hat{d}_r$ ) for  $k = 1, 2, \dots, t-2$ :

$$\begin{aligned} \hat{\mathbf{y}} &= fl\left(\mathbf{P}_0^T \mathbf{P}_1^T \mathbf{F}_1 \mathbf{P}_2^T \mathbf{F}_2 \cdots \mathbf{P}_{t-2}^T \mathbf{F}_{t-2} \mathbf{C}_{t-1} \hat{\mathbf{G}}_{t-2} \cdots \hat{\mathbf{G}}_2 \hat{\mathbf{G}}_1 \mathbf{G}_0 \mathbf{x}\right) \\ &= \mathbf{P}_0^T \mathbf{P}_1^T (\mathbf{F}_1 + \Delta \mathbf{F}_1) \cdots \mathbf{P}_{t-2}^T (\mathbf{F}_{t-2} + \Delta \mathbf{F}_{t-2}) (\mathbf{C}_{t-1} + \Delta \mathbf{C}_{t-1}) \\ &\quad (\hat{\mathbf{G}}_{t-2} + \Delta \hat{\mathbf{G}}_{t-2}) \cdots (\hat{\mathbf{G}}_2 + \Delta \hat{\mathbf{G}}_2) (\hat{\mathbf{G}}_1 + \Delta \hat{\mathbf{G}}_1) (\mathbf{G}_0 + \Delta \mathbf{G}_0) \mathbf{x} \end{aligned}$$

Each  $\mathbf{F}_k$  is formed containing a combination of matrices  $I_{\frac{n}{2^k}}$  and  $U_{\frac{n}{2^k}}$ . Using the fact that each row in  $\mathbf{F}_k$  has at most two non-zero entries with mostly ones per row:

$$|\Delta \mathbf{F}_k| \leq \gamma_2 |\mathbf{F}_k| \quad \text{for } k = 1, 2, \dots, t-2$$

Also each  $\mathbf{G}_k$  is formed containing a combination of matrices  $H_{\frac{n}{2^k}}$  and  $R_{\frac{n}{2^k}}$  except  $\mathbf{G}_0 = H_n$ . Using the fact that each row in  $\mathbf{G}_k$  has at most two non-zero entries per row:

$$|\Delta \mathbf{G}_0| \leq \gamma_2 |\mathbf{G}_0|, \quad |\Delta \hat{\mathbf{G}}_k| \leq \gamma_3 |\hat{\mathbf{G}}_k|, \quad \text{for } k = 1, 2, \dots, t-2$$

$\mathbf{C}_{t-1}$  is a block diagonal matrix containing  $C_2^{II}$  and  $C_2^{IV}$  hence

$$|\Delta \mathbf{C}_{t-1}| \leq \gamma_3 |\mathbf{C}_{t-1}|$$

Using direct call of computing trigonometric functions i.e. the view of (14),

$$\hat{\mathbf{G}}_k = \mathbf{G}_k + \Delta \mathbf{G}_k, \quad |\Delta \mathbf{G}_k| \leq \mu |\mathbf{G}_k|,$$

Thus, overall

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{P}_0^T \mathbf{P}_1^T (\mathbf{F}_1 + \Delta \mathbf{F}_1) \cdots \mathbf{P}_{t-2}^T (\mathbf{F}_{t-2} + \Delta \mathbf{F}_{t-2}) (\mathbf{C}_{t-1} + \Delta \mathbf{C}_{t-1}) \\ &\quad (\mathbf{G}_{t-2} + \Delta \mathbf{G}_{t-2}) \cdots (\mathbf{G}_2 + \Delta \mathbf{G}_2) (\mathbf{G}_1 + \Delta \mathbf{G}_1) (\mathbf{G}_0 + \Delta \mathbf{G}_0) \mathbf{x}, \end{aligned}$$

$$|\mathbf{E}_k| \leq (\mu + \gamma_3(1 + \mu)) |\mathbf{G}_k| \leq \gamma_5 |\mathbf{G}_k|$$

Hence

$$\begin{aligned} |\mathbf{y} - \hat{\mathbf{y}}| &\leq \left[ (1 + \gamma_2)^{t-1} (1 + \gamma_3) (1 + \gamma_5)^{t-2} - 1 \right] \mathbf{P}_0^T \mathbf{P}_1^T |\mathbf{F}_1| \cdots \\ &\quad \mathbf{P}_{t-2}^T |\mathbf{F}_{t-2}| |\mathbf{C}_{t-1}| |\mathbf{G}_{t-2}| |\mathbf{G}_{t-3}| \cdots |\mathbf{G}_0| |\mathbf{x}| \end{aligned}$$

where

$$\begin{aligned} (1 + \gamma_2)^{t-1} (1 + \gamma_3) (1 + \gamma_5)^{t-2} - 1 &\leq (1 + \gamma_2)^{t-1} (1 + \gamma_5)^{t-1} - 1 \\ &\leq (1 + \gamma_7)^{t-1} - 1 \leq \frac{\gamma_7(t-1)}{1 - \gamma_7(t-1)}. \end{aligned}$$

Since  $\mathbf{F}_k, \mathbf{C}_{t-1}, \mathbf{G}_k$  are orthogonal matrices,  $\|\mathbf{F}_k\|_2 = \|\mathbf{C}_{t-1}\|_2 = \|\mathbf{G}_k\|_2 = 1$ . By orthogonality of  $C_n^{II}$ ,  $\|\mathbf{y}\|_2 = \|\mathbf{x}\|_2$ . Hence

$$\|\mathbf{y} - \hat{\mathbf{y}}\|_2 \leq \frac{\gamma_7(t-1)}{1 - \gamma_7(t-1)} \|\mathbf{y}\|_2$$

□

**Corollary 4.2.**  $\mathbf{y} = C_n^{II} \mathbf{x}$  is forward and backward stable.

*Proof.* The above theorem says that radix 2 DCT-II yields a tiny forward error provided that  $\sin \frac{r\pi}{4n}$  and  $\cos \frac{r\pi}{4n}$  are computed stably. It immediately follows that the computation is backward stable because  $\hat{\mathbf{y}} = \mathbf{y} + \Delta \mathbf{y} = C_n^{II} \mathbf{x} + \Delta \mathbf{y}$  implies  $\hat{\mathbf{y}} = C_n^{II} (\mathbf{x} + \Delta \mathbf{x})$  with  $\frac{\|\Delta \mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \frac{\|\Delta \mathbf{y}\|_2}{\|\mathbf{y}\|_2}$ . If we form  $\mathbf{y} = C_n^{II} \mathbf{x}$  by using exact  $C_n^{II}$ , then  $|\mathbf{y} - \hat{\mathbf{y}}| \leq \gamma_n |C_n^{II}| |\mathbf{x}|$  so  $\|\mathbf{y} - \hat{\mathbf{y}}\|_2 \leq \gamma_n \|\mathbf{y}\|_2$ . As  $\mu$  is of order  $u$ , the  $C_n^{II}$  has an error bound smaller than that for usual multiplication by the same factor as the reduction in complexity of the method, so DCT-II is perfectly stable. □

The error bound of computing recursive DCT-IV algorithm can be derived as follows.

**Theorem 4.3.** Let  $\hat{\mathbf{y}} = fl(C_n^{IV} \mathbf{x})$ , where  $n = 2^t (t \geq 2)$ , be computed using the algorithms (2.2), (2.1), and assume that (14) holds, then

$$\frac{\|\mathbf{y} - \hat{\mathbf{y}}\|_2}{\|\mathbf{y}\|_2} \leq \frac{\gamma_7 t}{1 - \gamma_7 t}. \quad (16)$$

*Proof.* Using the algorithms (2.2), (2.1), and the computed matrices  $\widehat{\mathbf{W}}_k$  (in terms of the computed  $\widehat{d}_r$ ) for  $k = 0, 1, \dots, t-2$ :

$$\begin{aligned} \hat{\mathbf{y}} &= fl\left(\mathbf{P}_0^T \mathbf{U}_0 \mathbf{P}_1^T \mathbf{U}_1 \cdots \mathbf{P}_{t-2}^T \mathbf{U}_{t-2} \mathbf{C}_{t-1} \widehat{\mathbf{W}}_{t-2} \cdots \widehat{\mathbf{W}}_1 \widehat{\mathbf{W}}_0 \mathbf{x}\right) \\ &= \mathbf{P}_0^T (\mathbf{U}_0 + \Delta \mathbf{U}_0) \cdots \mathbf{P}_{t-2}^T (\mathbf{U}_{t-2} + \Delta \mathbf{U}_{t-2}) (\mathbf{C}_{t-1} + \Delta \mathbf{C}_{t-1}) \\ &\quad (\widehat{\mathbf{W}}_{t-2} + \Delta \widehat{\mathbf{W}}_{t-2}) \cdots (\widehat{\mathbf{W}}_1 + \Delta \widehat{\mathbf{W}}_1) (\widehat{\mathbf{W}}_0 + \Delta \widehat{\mathbf{W}}_0) \mathbf{x} \end{aligned}$$

Each  $\mathbf{U}_k$  is formed containing a combination of matrices  $I_{\frac{n}{2^k}}$  and  $U_{\frac{n}{2^k}}$  except  $\mathbf{U}_0 = U_n$ . Using the fact that each row in  $\mathbf{U}_k$  has at most two non-zero entries with mostly ones per row:

$$|\Delta \mathbf{U}_k| \leq \gamma_2 |\mathbf{U}_k| \text{ for } k = 0, 1, \dots, t-2$$

Also each  $\mathbf{W}_k$  is formed containing a combination of matrices  $H_{\frac{n}{2^k}}$  and  $R_{\frac{n}{2^k}}$  except  $\mathbf{W}_0 = R_n$ . Using the fact that each row in  $\mathbf{W}_k$  has at most two non-zero entries per row:

$$\begin{aligned} |\Delta \widehat{\mathbf{W}}_k| &\leq \gamma_3 |\widehat{\mathbf{W}}_k|, \\ \text{for } k &= 0, 1, \dots, t-2 \end{aligned}$$

$\mathbf{C}_{t-1}$  is a block diagonal matrix containing  $C_2^H$  and  $C_2^{IV}$  hence

$$|\Delta \mathbf{C}_{t-1}| \leq \gamma_3 |\mathbf{C}_{t-1}|$$

Using direct call of computing trigonometric functions i.e. the view of (14),

$$\widehat{\mathbf{W}}_k = \mathbf{W}_k + \Delta \mathbf{W}_k, \quad |\Delta \mathbf{W}_k| \leq \mu |\mathbf{W}_k|,$$

Thus, overall

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{P}_0^T (\mathbf{U}_0 + \Delta \mathbf{U}_0) \cdots \mathbf{P}_{t-2}^T (\mathbf{U}_{t-2} + \Delta \mathbf{U}_{t-2}) (\mathbf{C}_{t-1} + \Delta \mathbf{C}_{t-1}) \\ &\quad (\mathbf{W}_{t-2} + \mathbf{E}_{t-2}) \cdots (\mathbf{W}_1 + \mathbf{E}_1) (\mathbf{W}_0 + \mathbf{E}_0) \mathbf{x}, \\ |\mathbf{E}_k| &\leq (\mu + \gamma_3(1 + \mu)) |\mathbf{W}_k| \leq \gamma_5 |\mathbf{W}_k| \end{aligned}$$

Hence

$$\begin{aligned} \|\mathbf{y} - \hat{\mathbf{y}}\| &\leq \left[ (1 + \gamma_2)^{t-1} (1 + \gamma_3) (1 + \gamma_5)^{t-1} - 1 \right] \mathbf{P}_0^T |\mathbf{U}_0| \cdots \\ &\quad \mathbf{P}_{t-2}^T |\mathbf{U}_{t-2}| |\mathbf{C}_{t-1}| |\mathbf{W}_{t-2}| \cdots |\mathbf{W}_1| |\mathbf{W}_0| |\mathbf{x}| \end{aligned}$$

where

$$\begin{aligned} (1 + \gamma_2)^{t-1} (1 + \gamma_3) (1 + \gamma_5)^{t-1} - 1 &\leq (1 + \gamma_3) (1 + \gamma_7)^{t-1} - 1 \\ &\leq (1 + \gamma_7)^t - 1 \leq \frac{\gamma_7 t}{1 - \gamma_7 t}. \end{aligned}$$

Since  $\mathbf{U}_k, \mathbf{C}_{t-1}, \mathbf{W}_k$  are orthogonal matrices,  $\|\mathbf{U}_k\|_2 = \|\mathbf{C}_{t-1}\|_2 = \|\mathbf{W}_k\|_2 = 1$ . By orthogonality of  $C_n^{IV}$ ,  $\|\mathbf{y}\|_2 = \|\mathbf{x}\|_2$ . Hence

$$\|\mathbf{y} - \hat{\mathbf{y}}\|_2 \leq \frac{\gamma_7 t}{1 - \gamma_7 t} \|\mathbf{y}\|_2$$

□

**Corollary 4.4.**  $\mathbf{y} = C_n^{IV} \mathbf{x}$  is forward and backward stable.

*Proof.* The above theorem says that radix 2 DCT-IV yields a tiny forward error provided that  $\sin \frac{r\pi}{4n}$  and  $\cos \frac{r\pi}{4n}$  are computed stably. It immediately follows that the computation is backward stable because  $\hat{\mathbf{y}} = \mathbf{y} + \Delta \mathbf{y} = C_n^{IV} \mathbf{x} + \Delta \mathbf{y}$  implies  $\hat{\mathbf{y}} = C_n^{IV} (\mathbf{x} + \Delta \mathbf{x})$  with  $\frac{\|\Delta \mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \frac{\|\Delta \mathbf{y}\|_2}{\|\mathbf{y}\|_2}$ . If we form  $\mathbf{y} = C_n^{IV} \mathbf{x}$  by using exact  $C_n^{IV}$ , then  $|\mathbf{y} - \hat{\mathbf{y}}| \leq \gamma_n |C_n^{IV}| |\mathbf{x}|$  so  $\|\mathbf{y} - \hat{\mathbf{y}}\|_2 \leq \gamma_n \|\mathbf{y}\|_2$ . As  $\mu$  is of order  $u$ , the  $C_n^{IV}$  has an error bound smaller than that for usual multiplication by the same factor as the reduction in complexity of the method, so DCT-IV is perfectly stable. □

**Corollary 4.5.** Let  $\hat{\mathbf{y}} = fl(C_n^{III} \mathbf{x})$ , where  $n = 2^t (t \geq 2)$ , be computed using the algorithms (2.3), (2.2), (2.1), and assume that (14) holds, then

$$\frac{\|\mathbf{y} - \hat{\mathbf{y}}\|_2}{\|\mathbf{y}\|_2} \leq \frac{\gamma_7(t-1)}{1 - \gamma_7(t-1)}. \quad (17)$$

**Corollary 4.6.**  $\mathbf{y} = C_n^{III} \mathbf{x}$  is forward and backward stable.

Finally, the error bound for computing DCT-I algorithm, which runs recursively with DCT II-IV algorithms, can be derived as follows.

**Theorem 4.7.** Let  $\hat{\mathbf{y}} = fl(C_{n+1}^I \mathbf{x})$ , where  $n = 2^t (t \geq 2)$ , be computed using the algorithms (2.5), (2.3), (2.2), (2.1), and assume that (14) holds. Then

$$\frac{\|\mathbf{y} - \hat{\mathbf{y}}\|_2}{\|\mathbf{y}\|_2} \leq \frac{\gamma_7 t}{1 - \gamma_7 t}. \quad (18)$$

*Proof.* Using the algorithms (2.5), (2.3), (2.2), (2.1), and the computed matrices  $\widehat{\mathbf{B}}_k$  (in terms of the computed  $\widehat{d}_r$ ) for  $k = 2, 3, \dots, t-2$ :

$$\begin{aligned} \hat{\mathbf{y}} &= fl\left(\mathbf{A}_0 \mathbf{A}_1 \cdots \mathbf{A}_{t-2} \mathbf{C}_{t-1} \widehat{\mathbf{B}}_{t-2} \cdots \widehat{\mathbf{B}}_2 \mathbf{B}_1 \mathbf{B}_0 \mathbf{x}\right) \\ &= (\mathbf{A}_0 + \Delta \mathbf{A}_0) \cdots (\mathbf{A}_{t-2} + \Delta \mathbf{A}_{t-2}) (\mathbf{C}_{t-1} + \Delta \mathbf{C}_{t-1}) \\ &\quad (\widehat{\mathbf{B}}_{t-2} + \Delta \widehat{\mathbf{B}}_{t-2}) \cdots (\widehat{\mathbf{B}}_2 + \Delta \widehat{\mathbf{B}}_2) (\mathbf{B}_1 + \Delta \mathbf{B}_1) (\mathbf{B}_0 + \Delta \mathbf{B}_0) \mathbf{x} \end{aligned}$$

Each  $\mathbf{A}_k$  is formed containing a combination of matrices  $P_{\frac{n}{2^k}+1}^T, P_{\frac{n}{2^k}}^T, H_{\frac{n}{2^k}}^T$  and  $U_{\frac{n}{2^k}}$  except  $\mathbf{A}_0 = P_{n+1}^T$  and  $\mathbf{A}_1 = \text{blkdiag}\left(P_{\frac{n}{2}+1}^T, H_{\frac{n}{2}}^T\right)$ . Using the fact that each row in  $\mathbf{A}_k$  has at most two non-zero entries with mostly ones per row:

$$|\Delta \mathbf{A}_0| = 0, \quad |\Delta \mathbf{A}_k| \leq \gamma_2 |\mathbf{A}_k| \text{ for } k = 1, 2, \dots, t-2$$

Also each  $\mathbf{B}_k$  is formed containing a combination of matrices  $\check{H}_{\frac{n}{2^k}+1}, H_{\frac{n}{2^k}}^n, P_{\frac{n}{2^k}}^n$  and  $R_{\frac{n}{2^k}}$  except  $\mathbf{B}_0 = \check{H}_{n+1}$  and  $\mathbf{B}_1 = \text{blkdiag}\left(\check{H}_{\frac{n}{2}+1}, P_{\frac{n}{2}}^n\right)$ . Using the fact that each row in  $\mathbf{B}_k$  has at most two non-zero entries per row:

$$\begin{aligned} |\Delta \mathbf{B}_0| &\leq \gamma_2 |\mathbf{B}_0|, \quad |\Delta \mathbf{B}_1| \leq \gamma_2 |\mathbf{B}_1|, \quad |\Delta \widehat{\mathbf{B}}_k| \leq \gamma_3 |\widehat{\mathbf{B}}_k|, \\ \text{for } k &= 2, 3, \dots, t-2 \end{aligned}$$

$\mathbf{C}_{t-1}$  is a block diagonal matrix containing  $C_1^I$ ,  $C_2^I$ ,  $C_2^{III}$  and  $C_2^{IV}$  hence

$$|\Delta \mathbf{C}_{t-1}| \leq \gamma_3 |\mathbf{C}_{t-1}|$$

Using direct call of computing trigonometric functions i.e. the view of (14),

$$\hat{\mathbf{B}}_k = \mathbf{B}_k + \Delta \mathbf{B}_k, \quad |\Delta \mathbf{B}_k| \leq \mu |\mathbf{B}_k|,$$

Thus, overall

$$\hat{\mathbf{y}} = (\mathbf{A}_0 + \Delta \mathbf{A}_0) \cdots (\mathbf{A}_{t-2} + \Delta \mathbf{A}_{t-2}) (\mathbf{C}_{t-1} + \Delta \mathbf{C}_{t-1}) (\mathbf{B}_{t-2} + \Delta \mathbf{B}_{t-2}) \cdots (\mathbf{B}_2 + \Delta \mathbf{B}_2) (\mathbf{B}_1 + \Delta \mathbf{B}_1) (\mathbf{B}_0 + \Delta \mathbf{B}_0) \mathbf{x},$$

$$|\mathbf{E}_k| \leq (\mu + \gamma_3(1 + \mu)) |\mathbf{B}_k| \leq \gamma_5 |\mathbf{B}_k|$$

Hence

$$|\mathbf{y} - \hat{\mathbf{y}}| \leq \left[ (1 + \gamma_2)^t (1 + \gamma_3) (1 + \gamma_5)^{t-3} - 1 \right] |\mathbf{A}_0| |\mathbf{A}_1| \cdots |\mathbf{A}_{t-2}| |\mathbf{C}_{t-1}| |\mathbf{B}_{t-2}| |\mathbf{B}_{t-3}| \cdots |\mathbf{B}_0| |\mathbf{x}|$$

where

$$\begin{aligned} (1 + \gamma_2)^t (1 + \gamma_3) (1 + \gamma_5)^{t-3} - 1 &\leq (1 + \gamma_2)^t (1 + \gamma_5)^{t-2} - 1 \\ &\leq (1 + \gamma_7)^t - 1 \\ &\leq \frac{\gamma_7^t}{1 - \gamma_7^t}. \end{aligned}$$

Since  $\mathbf{A}_k, \mathbf{C}_{t-1}, \mathbf{B}_k$  are orthogonal matrices,  $\|\mathbf{A}_k\|_2 = \|\mathbf{C}_{t-1}\|_2 = \|\mathbf{B}_k\|_2 = 1$ . By orthogonality of  $\mathbf{C}_{n+1}^I$ ,  $\|\mathbf{y}\|_2 = \|\mathbf{x}\|_2$ . Hence

$$\|\mathbf{y} - \hat{\mathbf{y}}\|_2 \leq \frac{\gamma_7^t}{1 - \gamma_7^t} \|\mathbf{y}\|_2$$

□

**Corollary 4.8.**  $\mathbf{y} = \mathbf{C}_{n+1}^I \mathbf{x}$  is forward and backward stable.

*Proof.* The above theorem says that radix 2 DCT-I yields a tiny forward error provided that  $\sin \frac{\pi}{4n}$  and  $\cos \frac{\pi}{4n}$  are computed stably. It immediately follows that the computation is backward stable because  $\hat{\mathbf{y}} = \mathbf{y} + \Delta \mathbf{y} = \mathbf{C}_{n+1}^I \mathbf{x} + \Delta \mathbf{y}$  implies  $\hat{\mathbf{y}} = \mathbf{C}_{n+1}^I (\mathbf{x} + \Delta \mathbf{x})$  with  $\frac{\|\Delta \mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \frac{\|\Delta \mathbf{y}\|_2}{\|\mathbf{y}\|_2}$ . If we form  $\mathbf{y} = \mathbf{C}_{n+1}^I \mathbf{x}$  by using exact  $\mathbf{C}_{n+1}^I$ , then  $\|\mathbf{y} - \hat{\mathbf{y}}\|_2 \leq \gamma_{n+1} |\mathbf{C}_{n+1}^I| |\mathbf{x}|$  so  $\|\mathbf{y} - \hat{\mathbf{y}}\|_2 \leq \gamma_{n+1} \|\mathbf{y}\|_2$ . As  $\mu$  is of order  $u$ , the  $\mathbf{C}_{n+1}^I$  has an error bound smaller than that for usual multiplication by the same factor as the reduction in complexity of the method, so DCT-I is perfectly stable. □

## 5 IMAGE COMPRESSION RESULTS BASED ON DCT ALGORITHMS

Discretized images can be considered as matrices. To compress such images one can apply the quantization technique. In this section we use the quantization technique with the help of recursive DCT-II and DCT-IV algorithms to compress the Lena image of size  $512 \times 512$  pixels. At first, the image is discretized into  $8 \times 8$ ,  $16 \times 16$ , and  $32 \times 32$  transfer blocks. Next, using the recursive DCT-II and DCT-IV algorithms, 2D-DCTs are computed for each block.

The DCT-II and DCT-IV coefficients are then quantized by transforming absence of 93.75% of the DCT coefficients (93.75% of DCT-II and DCT-IV coefficients in each transfer block are set to zero). In each block, the inverse 2D DCT-II and DCT-IV coefficients are computed. Finally, putting each block back together into a single image leads to Figures 2 and 3.

Figure 2 shows images with discarded coefficients (except the top left 6.25% in each transfer block) in each transfer block, after applying DCT-II algorithm, and then running recursively with the DCT-IV algorithm.



Figure 2: (2a) Original Lena Image (2b) Reconstructed image with 93.75% discarded DCT-II coefficients in each  $8 \times 8$  transfer block (2c) Reconstructed image with 93.75% discarded DCT-II coefficients in each  $16 \times 16$  transfer block (2d) Reconstructed image with 93.75% discarded DCT-II coefficients in each  $32 \times 32$  transfer block

Figure 3 shows images with discarded coefficients (except the top left 6.25% in each transfer block) in each transfer block after applying DCT-IV algorithm and then running recursively with the DCT-II algorithm.

Comparing to Figures 2 and 3, the image reconstruction results corresponding to DCT-II algorithm are better than that of the DCT-IV algorithm. Though the quality of reconstructed images in Figures 2 and 3 are somewhat lost, those images are clearly recognizable even though 93.75%





Figure 3: (3a) Original Lena Image (3b) Reconstructed image with 93.75% discarded DCT-IV coefficients in each  $8 \times 8$  transfer block (3c) Reconstructed image with 93.75% discarded DCT-IV coefficients in each  $16 \times 16$  transfer block (3d) Reconstructed image with 93.75% discarded DCT-IV coefficients in each  $32 \times 32$  transfer block

of the DCT-II and DCT-IV coefficients are discarded in each transfer block.

## 6 SIGNAL FLOW GRAPHS FOR DCT ALGORITHMS

Signal flow graphs commonly represent the realization of systems such as electronic devices in electrical engineering, control theory, system engineering, theoretical computer science, etc. Simply put, the objective is to build a device to implement or realize an algorithm, using devices that implement the algebraic operations used in these recursive algorithms. These building blocks are shown next in Figure 4. This section presents signal flow graphs for 9-point DCT-I

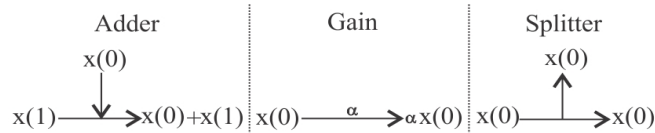


Figure 4: Signal flow graphs building blocks

and 8-point DCT II-IV algorithms via Figures 5, 6, 7, and 8. As shown in the flow graphs, in each graph signal flows from

the left to the right. These signal flow graphs are corresponding to the decimation-in-frequency algorithms. However one can convert these decimation-in-frequency DCT algorithms into decimation-in-time DCT algorithms. In each Figure (5, 6, 7, and 8),  $\epsilon := \frac{1}{\sqrt{2}}$ ,  $C_{i,j} := \cos \frac{i\pi}{2j}$ , and  $S_{i,j} = \sin \frac{i\pi}{2j}$ . As shown

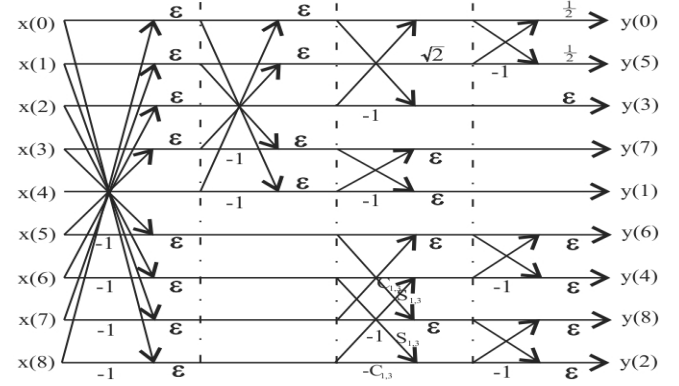


Figure 5: Flow graph for 9-point DCT-I algorithm

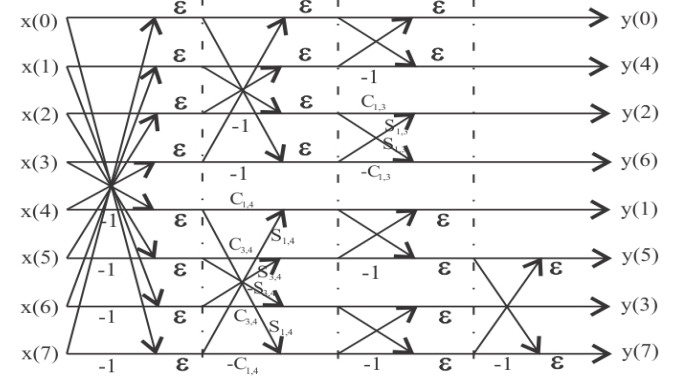


Figure 6: Flow graph for 8-point DCT-II algorithm

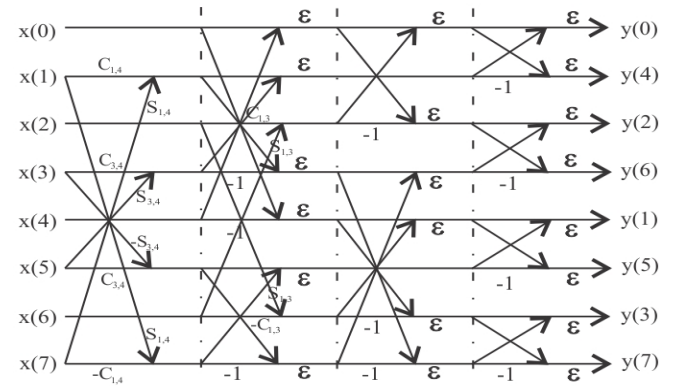


Figure 7: Flow graph for 8-point DCT-III algorithm

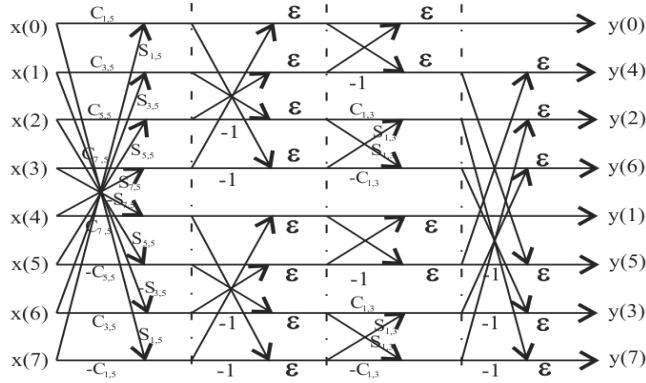


Figure 8: Flow graph for 8-point DCT-IV algorithm

in the Figures 6, 7, and 8, the input signals are in order:  $\mathbf{x} = \{x(0), x(1), \dots, x(7)\}$  and output signals are in bit-reversed order:  $\mathbf{y} = \{y(0), y(4), y(2), y(6), y(1), y(5), y(3), y(7)\}$ . In bit-reversed order, each output index is represented as a binary number and the indices' bits are reversed. Say for 8-point DCT II, the sequential order of the input indices' bits is  $\{000, 001, 010, 011, 100, 101, 110, 111\}$  then reversing these input signal bits yields  $\{000, 100, 010, 110, 001, 101, 011, 111\}$  which is the output signal.

## 7 CONCLUSION

This paper provided stable, completely recursive, radix-2 DCT-I and DCT-III algorithms having sparse, orthogonal and rotation/rotation-reflection matrices, defined solely via DCT I-IV algorithms. The arithmetic cost and error bounds of computing DCT I-IV algorithms are addressed. Using the recursive DCT-II and DCT-IV algorithms with the absence of 93.75% coefficients in each transfer block in 2D DCT-II and DCT-IV, one can reconstruct  $512 \times 512$  images without seriously affecting the quality. Signal flow graphs are presented for these solely based orthogonal factorization of DCT I-IV in decimation-of-frequency.

## REFERENCES

- [1] Ahmed, H., Natarajan, T., and Rao, K. R., 1974, "Discrete cosine transform", IEEE Trans. Comput., 23, 90-93.
- [2] Britanak, V., 2013, "New generalized conversion method of the MDCT and MDST coefficients in the frequency domain for arbitrary symmetric windowing function", Digital Signal Processing, 23, 1783-1797.
- [3] Britanak, V.; Yip, P. C.; Rao, K. R. 2007. Discrete Cosine and Sine Transforms: General Properties, Fast Algorithms and Integer Approximations, Academic Press, Great Britain.
- [4] Chakraborty, S., and Rao, K. R. 2012. "Fingerprint enhancement by directional filtering", In Proceeding of the 9th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, ECTICON, 2012, (Phetchaburi, Thailand, May 16-18), IEEE Xplore Digital Library, 1-4.
- [5] Chen, W. H., Smith, C.H., and Fralick, S., 1977, "A fast computational algorithm for the discrete cosine transform", IEEE Trans. Comm., 25, 1004-1009.
- [6] Fan, D., Meng, X., Wang, Y., Yang, X., Peng, X., He, W., Dong, G., and Chen, H., 2013, "Optical identity authentication scheme based on elliptic curve digital signature algorithm and phase retrieval algorithm", Applied Optics, 52, no. 23, 5645-52.
- [7] Han, J., Saxena, A., Melkote, V., and Rose, K., 2012, "Towards jointly optimal spatial prediction and adaptive transform in video/image coding", IEEE Transactions on Image Processing, 21, no. 4, 1874-1884.
- [8] Higham, N. J. 1961. Accuracy and Stability of Numerical Algorithms, SIAM Publications, Philadelphia, PA.
- [9] Jain, A. K., 1979, "A sinusoidal family of unitary transform", IEEE. Trans. Pattern Anal. Mach. Intell., PAMI-1, 356-365.
- [10] Jain, A. K., 1976, "A fast Karhunen-Loeve transform for a class of stochastic processes", IEEE. Trans. Commun., COM-24, 1023-1029.
- [11] Jain, P., Kumar, B., and Jain, S. B., 2009, "Unified recursive structure for forward and inverse modified DCT/DST/DHT", IETE Journal of Research, 55, no. 4, 180-191.
- [12] Kekre, H.B., Sarode, T. K., and Save, J. K., 2014, "Column Transform based Feature Generation for Classification of Image Database", International Journal of Application or Innovation in Engineering and Management, 3, no. 7, 172-181.
- [13] Kekre, H.B., Sarode, T., and Natu, P., 2014, "Performance Comparison of Hybrid Wavelet Transform Formed by Combination of Different Base Transforms with DCT on Image Compression", I.J. Image, Graphics and Signal Processing, 4, 39-45.
- [14] Kekre, H. B., and Solanki, J. K. , 1978, "Comparative performance of various trigonometric unitary transforms for transform image coding", Int. J. Electron., 44, 305-315.
- [15] Kim, D., and Rao, K. R., 2009, "2D-DST scheme for image mirroring and rotation", J. of Electronic Imaging, 17, no. 1., doi:10.1117/1.2885257.
- [16] Lee, M.H., Khan, M.H.A., Kim, K.J. , and Park, D., 2013, "A Fast Hybrid Jacket-Hadamard Matrix Based Diagonal Block-wise Transform", MITSUBISHI Electric Research Laboratories, TR2014-002.
- [17] Ma, J., Plonka, G., and Hussaini, M. Y. , 2012, "Compressive Video Sampling with Approximate Message Passing Decoding", IEEE Transactions on Circuits and Systems for Video Technology, 22, no. 9, 1354-1364.
- [18] Perera M., S., and Olshevsky, V., 2013, "Stable, Recursive and Fast Algorithms for DST having Orthogonal Factors", Journal of Coupled Systems Multiscale Dynamics, 1, 358-371.
- [19] Plonka, G., and Tasche, M., 2005 "Fast and Numerically stable algorithms for discrete cosine transforms", Linear Algebra and its Applications, 394, 309-345.
- [20] Potts, D., Steidl, G., and Tasche, M., 2002, "Numerical stability of fast trigonometric transforms - a worst case study", J. Concrete Appl. Math., 1, 1-36.
- [21] Puschel, M., and Moura, J. M. , 2003, "The algebraic approach to the discrete cosine and sine transforms and their fast algorithms", SIAM J. Comput., 32, 1280-1316.
- [22] Schreiber, U., 1986, Fast and Numerically stable trigonometric transforms, Thesis, University of Rostock, Germany.
- [23] Strang, G., 1986. Introduction to Applied Mathematics, Wellesley-Cambridge Press, MA.
- [24] Strang, G., 1999, "The Discrete Cosine Transform", SIAM Review, 41, 135-147.
- [25] Steidl, G., and Tasche, M., 1991, "A polynomial approach to fast algorithms for discrete Fourier-cosine and Fourier-sine transforms", Math. Comput., 56, 281-296.
- [26] Tasche, M., and Zeuner, H., 2000, "Roundoff error analysis for fast trigonometry transforms", in G. Anastassiou(Ed.), Handbook of Analytic-Computational Methods in Applied Mathematics, Chapman and Hall/CRC press, Boca Raton, 357-406.

- [27] Van Loan, C. 1992. Computational Frameworks for the Fast Fourier Transform, SIAM Publications, Philadelphia, PA.
- [28] Veerla, R., Zhang, Z., and Rao, K. R., 2012, "Advanced Image Coding and its Comparison with Various Still Image Codecs", American Journal of Signal Processing, 2, no. 5, 113-121.
- [29] Voronenko, Y., and Püschel, M., 2009, "Algebraic Signal Processing Theory: Cooley-Tukey Type Algorithms for Real DFTs", Transactions on Signal Processing, 57, no. 1, 1-19.
- [30] Wang, Z., 1984, "Fast algorithms for the discrete W transform and the discrete Fourier transform", IEEE Trans. Acoust. Speech Signal Process, 32, 803-816.
- [31] Wang, Z., and Hunt, B. R., 1983, "The discrete cosine transform-A new version", in Proc. Int. Conf. Acoust., Speech, Signal Processing, 1256 - 1259.
- [32] Yip, P., and Rao, K. R., 1980, "A fast computational algorithm for the discrete sine transform", IEEE Trans. Commun., 28, 304-307.